Constant Audibert et Camille CARRACO

# SAE : Régression sur des données réelles

## Notre démarche :

Le marché immobilier est un secteur dynamique et complexe, influencé par une multitude de facteurs économiques, sociaux et environnementaux. Dans le cadre de ce projet, nous nous intéressons aux ventes immobilières d'appartements et de maisons dans les Deux-Sèvres, couvrant l'année 2023 et le premier semestre 2024. L'objectif est de développer un modèle de prédiction du prix de vente des logements en utilisant un jeu de données (scindé en deux fichiers CSV : "train" et "test").

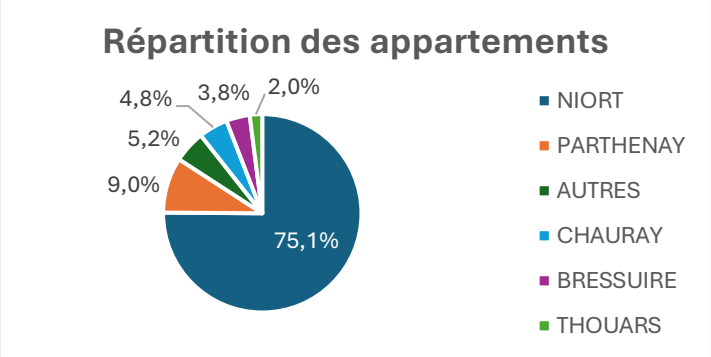
Le fichier "train" contient le prix de vente ainsi que diverses variables explicatives relatives à chaque logement, telles que la superficie, le nombre de pièces, l'année de construction, la localisation, et d'autres caractéristiques spécifiques. Le fichier "test", quant à lui, contient les mêmes variables explicatives, mais sans le prix de vente.

Notre démarche consiste donc à analyser les données, identifier les variables et les liens pertinents, puis élaborer un modèle de prédiction aussi fiable que possible. Ce modèle sera ensuite appliqué aux données du fichier "test" pour estimer les prix de vente des logements.

Notre objectif était de déterminer la valeur foncière. À notre disposition, nous avions plusieurs variables telles que la commune, la surface réelle, la surface du terrain, le nombre de pièces, etc.

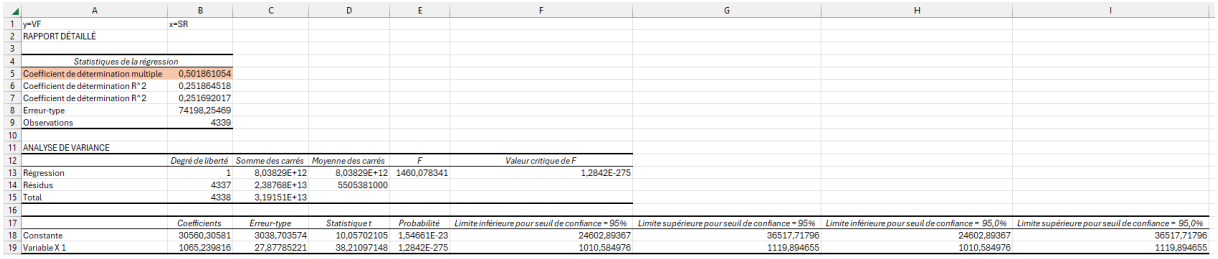
Nous avons tout d'abord exploré le jeu de données "test", cherchant à voir certaines relations entre les variables, comme la relation entre la valeur foncière et le nombre de pièces, la valeur foncière et la surface réelle, ou encore la valeur foncière et la surface du terrain. Nous nous sommes aussi intéressés aux variables qualitatives notamment la commune et le type du biens.

D'autre part, nous avons réalisé une analyse statistique descriptive qui rend compte du profil de notre jeu de données "train". On note que le jeu est composé à 91 % de maisons et à 9 % d'appartements. Nous remarquons aussi que trois quarts des appartements sont situés à Niort. Nous avons donc décidé d'isoler les maisons et



les appartements pour les analyser séparément.

Par la suite, nous avons cherché des variables qui pourraient expliquer la valeur foncière. Nous avons tracé des nuages de points, utilisé l'utilitaire de régression linéaire d'Excel et



| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | y=VF | x=SR | | | | | | | |
| 2 | RAPPORT DÉTAILLÉ | | | | | | | | |
| 3 | | | | | | | | | |
| 4 | *Statistiques de la régression* | | | | | | | | |
| 5 | Coefficient de détermination multiple | 0,501861054 | | | | | | | |
| 6 | Coefficient de détermination R^2 | 0,251864518 | | | | | | | |
| 7 | Coefficient de détermination R^2 | 0,251692017 | | | | | | | |
| 8 | Erreur-type | 74198,25469 | | | | | | | |
| 9 | Observations | 4339 | | | | | | | |
| 10 | | | | | | | | | |
| 11 | ANALYSE DE VARIANCE | | | | | | | | |
| 12 | | *Degré de liberté* | *Somme des carrés* | *Moyenne des carrés* | *F* | *Valeur critique de F* | | | |
| 13 | Régression | 1 | 8,03829E+12 | 8,03829E+12 | 1460,078341 | 1,2842E-275 | | | |
| 14 | Résidus | 4337 | 2,38768E+13 | 5505381000 | | | | | |
| 15 | Total | 4338 | 3,19151E+13 | | | | | | |
| 16 | | | | | | | | | |
| 17 | | *Coefficients* | *Erreur-type* | *Statistique t* | *Probabilité* | *Limite inférieure pour seuil de confiance = 95%* | *Limite supérieure pour seuil de confiance = 95%* | *Limite inférieure pour seuil de confiance = 95,0%* | *Limite supérieure pour seuil de confiance = 95,0%* |
| 18 | Constante | 30560,30581 | 3038,703574 | 10,05702105 | 1,54661E-23 | 24602,89367 | 36517,71796 | 24602,89367 | 36517,71796 |
| 19 | Variable X 1 | 1065,239816 | 27,87785221 | 38,21097148 | 1,2842E-275 | 1010,584976 | 1119,894655 | 1010,584976 | 1119,894655 |

testé des transformations logarithmiques et exponentielles, mais rien de vraiment significatif n'en a résulté.

| Commune | Prix moyen m**2 |
|---|---|
| THENEZAY | 240 |
| SAINT-POMPAIN | 471,0144928 |
| ARGENTONNAY | 511,9047619 |
| MAUZE-SUR-LE-MIGNON | 587,5 |
| BEAUVOIR-SUR-NIORT | 632,9113924 |
| MELLE | 686,2745098 |
| CELLES-SUR-BELLE | 805,2287582 |
| LES FORGES | 899,6872428 |
| EXIREUIL | 1072,327044 |
| ST MAIXENT L ECOLE | 1083,472222 |
| THOUARS | 1333,025424 |
| PARTHENAY | 1446,066947 |
| BRESSUIRE | 1525,738699 |
| FRONTENAY-ROHAN-ROHAN | 1704,545455 |
| LA MOTHE SAINT HERAY | 1847,826087 |
| NIORT | 1895,199878 |
| CHEF-BOUTONNE | 2072,5 |
| CHAURAY | 2089,440228 |
| VOUILLE | 3446,324265 |

*Tableau 1 Prix moyen m^2 (appartement)*

Après des échanges avec nos collègues et l'exploration de modèles d'IA, nous avons affiné notre compréhension du jeu de données. C'est alors qu'il nous est venu à l'idée de calculer le prix au $m^2$. Nous avons ensuite calculé le prix moyen par commune et formé un tableau de référence avec le prix moyen au $m^2$ pour chaque commune (recherche effectuée). Cela nous permet ainsi de nous faire une idée globale du marché immobilier des Deux-Sèvres en fonction de la commune.

À l'aide du prix moyen au $m^2$ pour chaque commune, on pourrait simplement multiplier la surface réelle par le prix moyen au $m^2$. Cependant, cela donnerait une estimation brute, car cette méthode ne tient pas compte de la relation plus complexe entre ces deux variables et la valeur foncière.

# Notre modèle :

Ainsi nous avons choisi de retenir un modèle à régression linéaire :

$$VF = \beta 0 + \beta 1 \cdot SR + \beta 2 \cdot PM2C + \beta 3 \cdot (SR \times PM2C)$$

Avec le jeu de données, R calcule les coefficients appropriés (fonction lm).

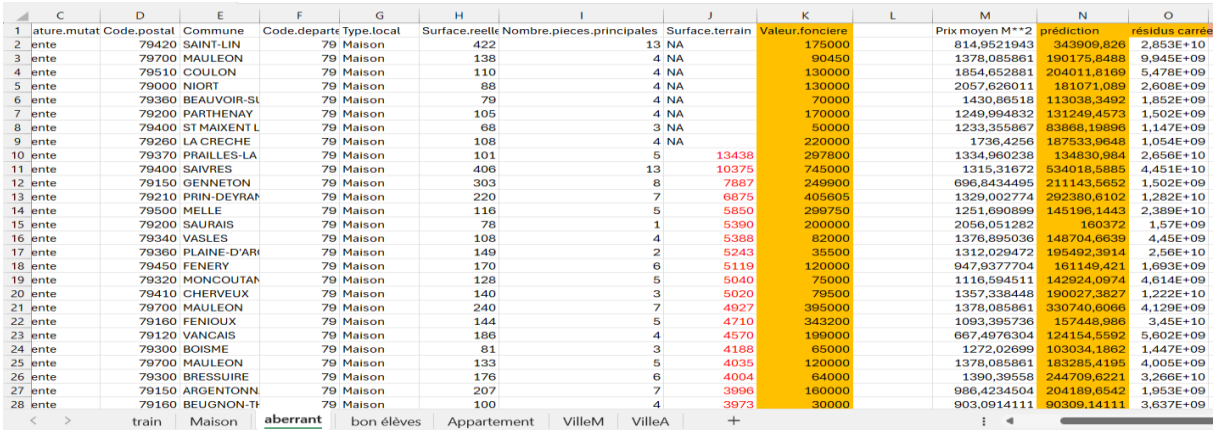
**$\beta_0$ (Intercept)** : Cela représente la valeur de départ . Soit, la valeur d'un bien si la surface et le prix moyen au $m^2$ étaient nuls).

**$\beta_1$** : C'est le coefficient associé à la surface réelle. Il représente l'augmentation de la valeur foncière pour chaque mètre carré supplémentaire de surface bâtie, lorsque le prix moyen au $m^2$ reste constant.

**$\beta_2$** : C'est le coefficient associé au prix moyen au $m^2$ par commune. Il indique l'effet du prix moyen par $m^2$ de la commune sur la valeur foncière, lorsque la surface bâtie est maintenue constante.

**$\beta_3$** : C'est le coefficient associé à l'interaction surface réelle × prix moyen au $m^2$ par commune. Il montre comment l'effet de la surface bâtie sur la valeur foncière change en fonction du prix moyen au $m^2$ de la commune.

C'est le modèle le plus performant que nous avons réussi à produire. Celui-ci prédit assez bien les valeurs foncières du fichier "train", bien que certaines valeurs aberrantes fassent grimper le $R^2$. Nous avons essayé d'améliorer ce modèle en nous concentrant sur les pires prédictions (carré des résidus > 1 000 000 000) et les meilleures prédictions (carré



| | C | D | E | F | G | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ature.mutat | Code.postal | Commune | Code.departe | Type.local | Surface.reelle | Nombre.pieces.principales | Surface.terrain | Valeur.fonciere | | Prix moyen M**2 | prédiction | résidus carrée |
| 2 | ente | 79420 | SAINT-LIN | 79 | Maison | 422 | 13 | NA | 175000 | | 814,9521943 | 343909,826 | 2,853E+10 |
| 3 | ente | 79700 | MAULEON | 79 | Maison | 138 | 4 | NA | 90450 | | 1378,085861 | 190175,8488 | 9,945E+09 |
| 4 | ente | 79510 | COULON | 79 | Maison | 110 | 4 | NA | 130000 | | 1854,652881 | 204011,8169 | 5,478E+09 |
| 5 | ente | 79000 | NIORT | 79 | Maison | 88 | 4 | NA | 130000 | | 2057,626011 | 181071,089 | 2,608E+09 |
| 6 | ente | 79360 | BEAUVOIR-SU | 79 | Maison | 79 | 4 | NA | 70000 | | 1430,86518 | 113038,3492 | 1,852E+09 |
| 7 | ente | 79200 | PARTHENAY | 79 | Maison | 105 | 4 | NA | 170000 | | 1249,994832 | 131249,4573 | 1,502E+09 |
| 8 | ente | 79400 | ST MAIXENT L | 79 | Maison | 68 | 3 | NA | 50000 | | 1233,355867 | 83868,19896 | 1,147E+09 |
| 9 | ente | 79260 | LA CRECHE | 79 | Maison | 108 | 4 | NA | 220000 | | 1736,4256 | 187533,9648 | 1,054E+09 |
| 10 | ente | 79370 | PRAILLES-LA | 79 | Maison | 101 | 5 | 13438 | 297800 | | 1334,960238 | 134830,984 | 2,656E+10 |
| 11 | ente | 79400 | SAIVRES | 79 | Maison | 406 | 13 | 10375 | 745000 | | 1315,31672 | 534018,5885 | 4,451E+10 |
| 12 | ente | 79150 | GENNETON | 79 | Maison | 303 | 8 | 7887 | 249900 | | 696,8434495 | 211143,5652 | 1,502E+09 |
| 13 | ente | 79210 | PRIN-DEYRAN | 79 | Maison | 220 | 7 | 6875 | 405605 | | 1329,002774 | 292380,6102 | 1,282E+10 |
| 14 | ente | 79500 | MELLE | 79 | Maison | 116 | 5 | 5850 | 299750 | | 1251,690899 | 145196,1443 | 2,389E+10 |
| 15 | ente | 79200 | SAURAIS | 79 | Maison | 78 | 1 | 5390 | 200000 | | 2056,051282 | 160372 | 1,57E+09 |
| 16 | ente | 79340 | VASLES | 79 | Maison | 108 | 4 | 5388 | 82000 | | 1376,895036 | 148704,6639 | 4,45E+09 |
| 17 | ente | 79360 | PLAINE-D'AR | 79 | Maison | 149 | 2 | 5243 | 35500 | | 1312,029472 | 195492,3914 | 2,56E+10 |
| 18 | ente | 79450 | FENERY | 79 | Maison | 170 | 6 | 5119 | 120000 | | 947,9377704 | 161149,421 | 1,693E+09 |
| 19 | ente | 79320 | MONCOUTAN | 79 | Maison | 128 | 5 | 5040 | 75000 | | 1116,594511 | 142924,0974 | 4,614E+09 |
| 20 | ente | 79410 | CHERVEUX | 79 | Maison | 140 | 3 | 5020 | 79500 | | 1357,338448 | 190027,3827 | 1,222E+10 |
| 21 | ente | 79700 | MAULEON | 79 | Maison | 240 | 7 | 4927 | 395000 | | 1378,085861 | 330740,6066 | 4,129E+09 |
| 22 | ente | 79160 | FENIOUX | 79 | Maison | 144 | 5 | 4710 | 343200 | | 1093,395736 | 157448,986 | 3,45E+10 |
| 23 | ente | 79120 | VANCAIS | 79 | Maison | 186 | 4 | 4570 | 199000 | | 667,4976304 | 124154,5592 | 5,602E+09 |
| 24 | ente | 79300 | BOISME | 79 | Maison | 81 | 3 | 4188 | 65000 | | 1272,02699 | 103034,1862 | 1,447E+09 |
| 25 | ente | 79700 | MAULEON | 79 | Maison | 133 | 5 | 4035 | 120000 | | 1378,085861 | 183285,4195 | 4,005E+09 |
| 26 | ente | 79300 | BRESSUIRE | 79 | Maison | 176 | 6 | 4004 | 64000 | | 1390,39558 | 244709,6221 | 3,266E+10 |
| 27 | ente | 79150 | ARGENTONN | 79 | Maison | 207 | 7 | 3996 | 160000 | | 986,4234504 | 204189,6542 | 1,953E+09 |
| 28 | ente | 79160 | BEUGNON-TH | 79 | Maison | 100 | 4 | 3973 | 30000 | | 903,0914111 | 90309,14111 | 3,637E+09 |

train | Maison | aberrant | bon élèves | Appartement | VilleM | VilleA

des résidus proche de 0). Nous avons essayé de comprendre les raisons pour lesquelles certaines valeurs foncières n'étaient pas prédites correctement par notre modèle en analysant les caractéristiques des biens. Cependant, nous n'avons tiré aucune conclusion significative de cette analyse, mis à part la surface du terrain (pour les maisons), qui pourrait expliquer un certain écart entre la prédiction et la valeur foncière réelle.

## Conclusion et ressenti :

En somme, ce travail nous a permis de mobiliser nos connaissances en régression linéaire à un contexte concret (Prédire des valeurs foncières).

Au début du projet nous étions globalement perdus, une question tournait en boucle.

“Comment prédire la valeur foncière efficacement ? ». Mais nous voulions aller trop vite. Trouvez un modèle proche de la perfection rapidement, sans réellement comprendre comment est constitué notre jeu de données. Nous faisions fausse route.

Ajouter à cela, nous ne savions pas comment exploiter les variables qualitatives dans le cadre d’une régression linéaire. C’était nouveau.